California Airbnb: Data Modeling, Analysis, and Visualization

Project Report

Data - 225 (Database Systems for Analytics)

Abinaya Seshadre, Megha Goushal, Nikhila Gunda, Vinupriya Sanjay Kumar, Vraj Bharatkumar Mistry

Abstract— Since the outset of Airbnb in 2008, it has grown significantly from a small online website, offering short-term bed and breakfast to a leading online marketplace that connects people who want to rent out their place and people who are looking for accommodations in that locality. Airbnb, currently, has its presence in more than 220 countries and 100,000 cities worldwide. This has caused a radical change in the hospitality industry. Airbnb offers a wide variety of rental options from a single room, cabins, cottages, entire homes, apartments, etc to be booked on their platform.

California is one of the most attractive tourist destinations in the US and Airbnb has helped travelers get a personalized and unique accommodation experience here. In this project, we will explore the Airbnb dataset to understand the rental perspectives in California state. We will use the datasets available to the public from the 'Inside Airbnb' website. We study the different factors that have influenced the Airbnb listings in multiple cities and compare the trends and patterns in customer booking. The key is to get a better understanding of the business and reveal some interesting insights through our data analysis.

Index Terms— H.2 Database Management, H.2.0.b Database design, modeling and management, H.2.0.c Query design and implementation languages, H.2.8 Database Applications

1 Introduction

THERE are over 38300 Airbnb listings in California as of February 2021, which approximates to around 4 houses being rented per square mile. Airbnb has seen an exponential increase in the number of listings in California each year and has gained a lot of popularity. By analyzing the number of listings and occupancy rates, we can understand demand rates and also provide metrics for people who would like to make an investment in Airbnb and rent out their properties. Previously collected data and reviews help understand how the occupancy rate can be increased.

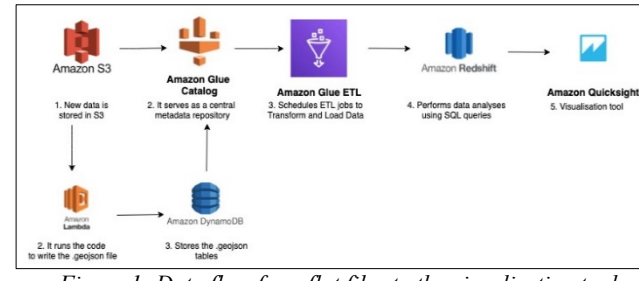

Figure 1: Data flow from flat files to the visualization tool

2 Methodology

During our initial steps, we went through the Airbnb dataset and analyzed the various column data available. We understood the metadata and gradually decided on which among the available datasets are required to achieve our goals. Also, data were available for almost all major cities of the world where Airbnb is available but being in the state of California with our university located in Santa Clara County with Airbnb presence, we decided to confine with the major cities in California state.

The below figure shows the process we undertook and the tools we utilized to ingest the data into our data warehouse and analytics and visualizations on the data.

2.1 Data Sources

The dataset used for our analysis was taken from the Inside Airbnb website. The dataset consisted of flat files having details regarding the listings, user reviews, and geojson files indicating the location of the listings in each neighborhood within the cities.

We downloaded the datasets for only the chosen counties of California and continued to ingest them into our chosen processing tools for further steps on the data, involving cleansing, loading into our data warehouse, and analysis.

2.2 Tools

Our team wanted to get exposed to cloud-based tools as those were not part of our database coursework or homework. Hence, we decided to explore, learn, and adopt AWS for our project. The following services of AWS were used for our project.

2.2.1 Amazon Simple Storage Service (S3)

As the name suggests, the Amazon Simple Storage Service acts as a file storage system and provides storage space for various

data and objects. S3 acts as a bucket that stores all our data and files. S3 has a scalable infrastructure and is a public resource in the cloud provided by Amazon Web Services.
The advantage of using an S3 bucket is that all the data is at one location and can be accessed by any of the services and tools provided by AWS. Also, the scalability factor makes it more beneficial for the implementation of projects.

2.2.2 Amazon Glue
Amazon Glue is a serverless processing and computing service provided by AWS. It is useful for data integration and the formation of metadata for table data. It acts as a crawler that crawls through the data to detect the schema and composes the metadata for the data, after which data is loaded into the chosen target.

Data can be pre-processed using Glue. Glue provides a data catalog, which could be used to analyze and understand our data better. It is an Extract Transform Load (ETL) and event-driven service to build our data in a better shape. The data is ingested into Glue with the help of a crawler and a job.

2.2.3 Amazon Redshift
Amazon Redshift is a fully managed, cloud-based data warehouse provided by AWS. It is SQL compatible and SQL queries are used to perform DDL and DML operations on the tables and the large datasets. Redshift can store and operate (process joins) on even petabyte-sized data.

2.2.4 Amazon DynamoDB
DynamoDb is a key-value NoSQL database cloud service. It supports both document-based and key-value-based data. It is highly scalable, fully managed with data security, delivers high performance, is reliable, with no limit to dataset size, flexible, and cost-efficient.

2.2.5 Amazon QuickSight
Amazon QuickSight is AWS's visualization tool. It is highly scalable and consists of business intelligence and machine learning capabilities built in to easily create insights, visuals, and dashboards for our applications, data models, and projects.

QuickSight provides access to load data from many sources like S3, Redshift which is AWS services, and sources external to AWS like PostgresSQL, Oracle, Teradata, and so on. It maintains a local dataset using spice. From the data available in spice, we can create any type of analysis using the UI controls. We can also create graphs and charts using our custom SQL queries.

3 Methods Incorporated
In this section, the methods adopted in the usage of our chosen tools are elaborated. The following are the various phases we followed right from the data acquisition stage to the business report stage.

3.1 Data Storage
After collecting all the required data files from the data sources, we uploaded the raw data into Amazon S3. S3 can be considered as the initial step in the AWS pipeline from which we can access and load files to other AWS resources.

We created a bucket for our project and then uploaded the required CSV files into that bucket. All our datasets were stored in a single bucket named 'airbnbproj'. After loading the data into the bucket, it consisted of all the files and folders for each county. For the implementation of the data model and design, the data was accessed from S3.

We also used AWS Redshift to store the data in the form of tables. Alternatively, in the case of semi-structured data, like the geojson files, NoSQL databases (DynamoDB) were used to extract and store the data from S3.

3.2 Data Processing
In this phase, we used the AWS Glue to perform ETL operations including data cleaning.

Firstly, a crawler is created which has the source file extracted from the S3 bucket to get the data. Then, a corresponding job is created for the crawler to crawl through the data to detect the schema, metadata types and provide a data catalog for further processing. It then builds and gives a 'PySpark' script to clean, process, and transform the data according to our requirements.

A crawler with a corresponding job for each county was created to perform the ETL services on our data.

Additionally, the data was cleaned by removing the null values, removing duplicates, formatting certain text fields to lowercase for easy analysis, and taking care of the date format for the date fields. The Glue script runs through the data, transforms it, and stores the table in Redshift.

3.3 Data Loading
In this phase, we load the data into Redshift from two different sources. The first source uses AWS Glue job to load data into Redshift by specifying the connection details like the database name, username, password, Identity and Access Management (IAM) Role associated with Redshift in the Glue job. The data is ingested through the Glue job as tables into a particular database in Redshift.

For the second source, we created a python script to load geoJSON data into the DynamoDB tables. From DynamoDB, we loaded the tables directly into Redshift by creating table schema and writing and executing copy commands on the Redshift query editor to copy the data from DynamoDB tables.

Redshift consists of a cluster with the corresponding Identity and access management (IAM) role which gives the specific access permissions to set up the environment for the data warehouse. Once the cluster is started and the database is connected to the query editor, operations can be performed to analyze data. The Glue job created all the tables in this database and now SQL queries can be written to understand data. The output is shown on the same console for the queries which can be further visualized using other visualization tools. We can also preview

the table data and the schema for our reference.

Our dataset had the geojson files which were stored in DynamoDB. The geojson files consisted of the information regarding the latitudes and longitudes for all the Airbnb listings in different counties of California. The data was first uploaded and stored in the S3 bucket and later was ingested into DynomoDb using the AWS Lambda function. The Python script of the Lambda function was modified according to our requirement to perform operations to write the tables in DynamoDB. These tables were later transferred to Redshift using a Glue job.

3.4 Data Modelling

The following is our data model, and it included the following entities:

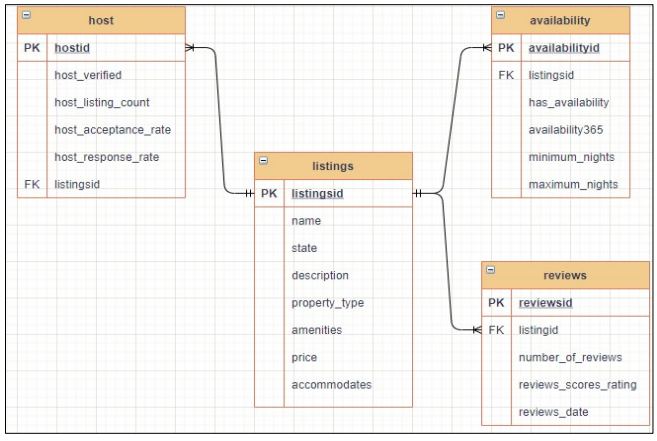

Figure 2: Data Model

4.4.1 Listings

This entity has all the details regarding the Airbnb listings including the property name, description, neighborhood, amenities, price, and how many people it can accommodate.

4.4.2 Host

This entity has all the details regarding the Airbnb property hosts and their response rates. This data is included to help check if attributes relating to the hosts affect Airbnb business, during our analysis stage.

4.4.3 Availability

This entity has all the details regarding the Airbnb property availability. This data is included to help check if attributes relating to availability affect the Airbnb business, during our analysis stage.

4.4.4 Reviews

This entity has all the details regarding the Airbnb property reviews. This data is included to help check if attributes relating to the reviews affect the Airbnb business, during our analysis stage.

3.5 Data Analytics

We used SQL queries extensively in our data warehouse which helped us arrive at interesting and useful insights from our data. We were able to analyze and look at data on applying various filter conditions and arrive at various conclusions, some of which are shared below.

After created the redshift cluster with all the required permissions using the IAM role to create the environment for our database and loading the necessary data, the connection was made to the database and SQL queries were performed on each county tables to determine the number of listings in different neighborhoods, reviews per listings, ratings for different listings, the average price for each county, etc.

The SQL queries were used to join different tables using the union operation. A different set of queries were written for each county listing table to analyze the data better county-wise. The query results were further used for visualization.

3.5.1 SQL Queries

The data was analyzed using SQL queries. A sample of our data analysis using SQL queries in Redshift is given below.

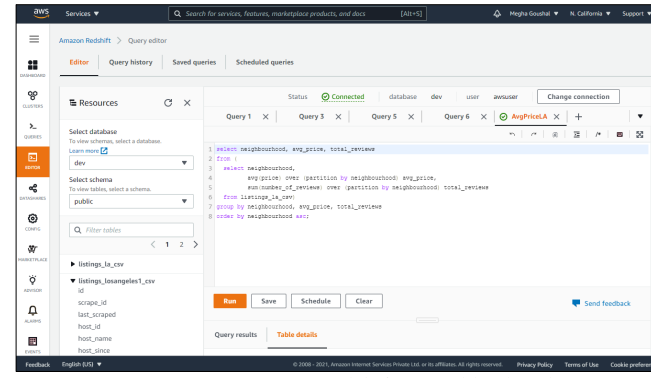

Figure 3: SQL Query in AWS Redshift - Neighborhood, price, reviews correlation

Rows returned (258)

Export

Search rows

1 2 3 4 5 6 7 ... 26

neighbourhood	avg_price	total_reviews
Acton	94	53
Adams-Normandie	72	1479
Agoura Hills	243	1188
Agua Dulce	229	568
Alhambra	104	10863
Alondra Park	133	596
Altadena	157	9010
Angeles Crest	206	282
Arcadia	142	2575
Arleta	58	229

Figure 4: Query result table - Neighborhood, price, reviews the correlation

The above query gives the count of reviews in each neighborhood of the chosen county, in this case, Los Angeles, and gives the average price of each neighborhood. Users and analysts can easily know and distinguish expensive and not-so-expensive areas of the listings using this query and can further analyze their validity, reasons, and business use.

```
select review_year, county, availability, count(availability)
from (
  select id, split_part(first_review,'-',1) as review_year, 'Los Angeles' as county, 'Less than 30' as availability
  from listings_losangeles1_csv
  where availability_365 < 30
  union all
  select id, split_part(first_review,'-',1) as review_year, 'Los Angeles' as county, 'Less than 60' as availability
  from listings_losangeles1_csv
  where availability_365 < 60
  union all
  select id, split_part(first_review,'-',1) as review_year, 'Los Angeles' as county,'More than 100' as availability
  from listings_losangeles1_csv
  where availability_365 > 100
  union all
  select id, split_part(first_review,'-',1) as review_year, 'Los Angeles' as county, 'More than 200' as availability
  from listings_losangeles1_csv
```

Run | Save | Schedule | Clear | Send feedback

Figure 5: Availability Query in Redshift Query Editor

review_year	county	availability	count
2017	Los Angeles	Less than 30	850
2017	Los Angeles	Less than 60	948
2017	Los Angeles	More than 100	1137
2017	Los Angeles	More than 200	773
2017	Los Angeles	More than 300	543
2017	San Francisco	Less than 30	153
2017	San Francisco	Less than 60	195
2017	San Francisco	More than 100	233
2017	San Francisco	More than 200	131
2017	San Francisco	More than 300	67

Figure 6: Availability Query Result Table

Many such queries were executed in the Redshift console and results were used to draw useful insights and conclusions and visualizations.

4 Data Visualization

4.1 Tools

After satisfied with our data analysis step, we were eager to have a look at how our data looked on maps, graphs, and charts for better comprehension and to make out trends and patterns. We turned to both MS Excel charts and AWS QuickSight for visualization.

As an initial step, visualization using Microsoft Excel was undertaken as it is a tool familiar to all of us. We generated various pivot charts and tried to capture information from each of the charts.

We connected QuickSight to our Redshift cluster to analyze and visualize the queries and tables existing in our Redshift database. Bar graphs, charts were used for visualizing the total number of Airbnb listings, their average prices, reviews, and so on for different counties. Geospatial maps were used to visualize the latitudes and longitudes for all counties.

4.2 Graphs and Tables

The number of listings in a particular county, type of listings that users of Airbnb prefer, the popularity of Airbnb among the people, price ranges of the listings across the counties is few parameters that can be visualized either separately or combined to capture useful information depicted that can be of great help to both business stakeholders and the public.

In this section, we show a few of the visualizations that were done and their output graphs, maps, and charts.

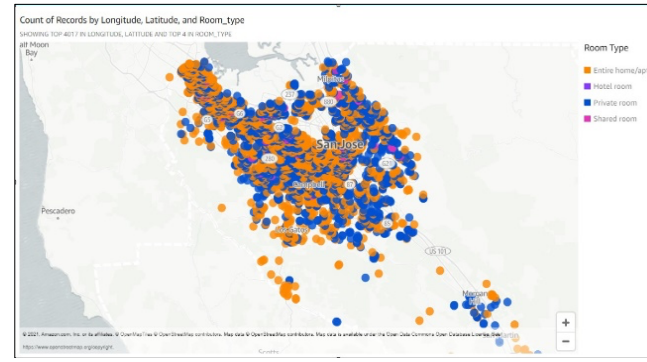

Figure 7: Map Visualization of Airbnb listings and their types in Santa Clara County, CA

The above map illustrates the listings of Airbnb in the city of Santa Clara. Further, the classification of each of the listings which are based on the type of accommodation such as Entire home, Private room, Shared room, Apartment is also shown.

This kind of visualization helps to ponder and answer the following questions:

1. Are the listings clustered to a single neighborhood within the county?
2. If the listings are clustered, why might it be so?
3. Are all listing types equally available?
4. What type of listing is more available, and will this match with people's requirements?
5. Why is a listing more in this area? Are there specific amenities or host attributes that should be uncovered that will help the business?

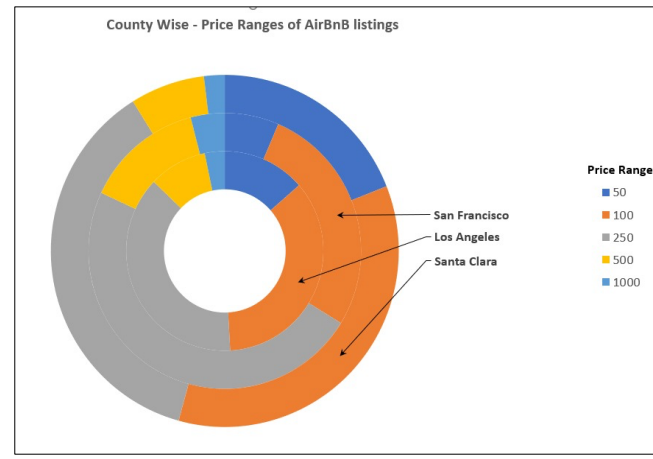

Figure 8: Price ranges of listings across the selected counties in California

The above chart correlates the price of the listings to the number in each price range. The price data in our datasets had values ranging from $5 to as much as $7000. So, SQL query was first used to bring the prices to fit into price ranges starting from $50 and incrementing by $50 for every step. There were comparatively very few listings below $50 and above $1000 and hence, such properties were not marked with distinct price ranges and were instead included in the below $50, above $100 ranges, respectively.

The pie chart depicts the maximum number of Airbnb listings are within the $50 - $250 price ranges. This insight can help the financial and business analysts at Airbnb predict their future revenues. This also helps Airbnb hosts during the process of fixing a price for their property. This can also help future users budget their trips and answer whether Airbnbs is a viable option compared to hotel stays.

If this were done as a map, as previously depicted, instead of a chart, we would know where the low-priced, moderately priced, and high-priced listings are located across the various chosen counties.

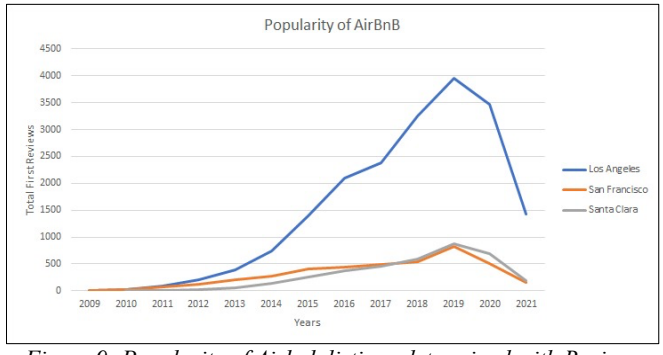

Figure 9: Popularity of Airbnb listings determined with Review Data

The above visualization was created with the year of the first review for each listing on the X-axis and the total number of reviews on the Y-axis. This shows Airbnb was not so popular in the early 2010s and people started using it more from around 2013 to 2014 timeframe. Significant rise in reviews started from 2014 onwards and grew drastically until early 2019. The drop in reviews from 2019 to 2020 time could be due to the Covid pandemic.

The visual also clearly shows Airbnbs are highly popular and highly utilized In Los Angeles county than the other chosen counties of California.

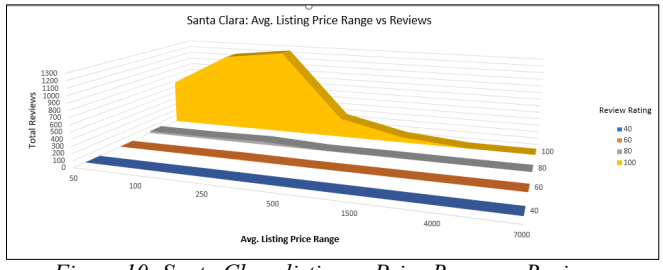

Figure 10: Santa Clara listings - Price Range vs Reviews

The data for the above visual were collected based on highly accurate reviews and the average rating from the latest 5 reviews for each listing. Total reviews were taken on Y-axis and average listing price ranges were taken on X-axis and review ratings were taken on Z-axis. Based on the 3-dimensional area chart, we could see Airbnb listings have a top rating of 100 for almost all their listings and a smaller number of listings around a rating of 80. Poor reviews were extremely low for their listings. Also, we understand from this chart, the listings in the price range of $100 to $250 have the highest ratings as well as the greatest number of ratings. If we narrow it down further, the listings in the price range of $200 to $250 are probably the best quality ones.

5 Deliverables and Milestones

The following table shows the timeline we followed for our project:

Task	Duration
Data Sources	Mar 10 – Mar 16
Data Storage	Mar 16 – Mar 23
Data Cleansing	Mar 23- Mar 31
Data Loading	Mar 31 – Apr 6
Data Modelling	Apr 6 – Apr 15
Data Analyzing	Apr 15 – Apr 25
Data Visualization	Apr 25 – May 1
Testing	May 1 – May 7
Report Draft	May 1 – May 7
Final Report	May 7 – May 11

6 Team

The team consisted of 5 students as listed below.

1. Abinaya Seshadre
2. Megha Goushal
3. Nikhila Gunda
4. Vinupriya Sanjay Kumar
5. Vraj Bharatkumar Mistry

All of us participated in all the tasks and practiced pair programming. We followed agile methodology and met every day via Zoom meetings and constantly updated, supported each other, and had timely deliverables to help complete all tasks.

7 Technical Difficulties Faced

As a part of our project development, we did face challenges. Listed are the technical difficulties we faced and tried to resolve:

1. Eliminating null values using the python script was a challenge initially and we could not achieve that
2. While creating the AWS Glue data catalog using crawler and running the glue jobs, we were not able to run the jobs correctly. We needed to find a way to debug the logs. We finally figured out that we had to create a VPC and add a classifier for the CSV in the Glue job.

3. Our data has geojson files. Initially, when creating tables using the lambda function in Dynamo DB, the table got created in the form of a complex nested JSON. We could not create tables in Redshift from the JSON. Then, we used Python script to connect to DynamoDB and then read the JSON files and created the tables.
4. While we were trying to connect to QuickSight using Redshift, we could not access the tables in Redshift. We had to create a security group with inbound rules to connect establish the connection between Redshift and quicksight as they both were in different regions.

8 Conclusion

It is immensely helpful for any business or user to make use of unused but accumulated data for analysis in the same way it was undertaken for this project. Gone are the days when companies as well as individuals had to second guess and take chances on their conclusions and viewpoints. This is an era of high technological advances in which even individuals sitting at home can wield power from their private devices and single-handedly conduct extensive analysis and visualization of data of their choice and come up with interesting insights that might help them personally.

The analysis and visualizations here can be further enhanced with the inclusion of even more parameters. Machine learning could be incorporated for the prediction of property prices, business growth, and unravel and be prepared for new opportunities.

9 Appendices

Acknowledgment

The entire team would like to extend our heartfelt gratitude to our professor, Dr. Vishnu S. Pendyala for giving us this opportunity to bring into practical use the lessons we learned from his lectures and guiding us throughout the entire project. Due to his constant motivation, we strived to do our best in exploring and learning new tools, methodologies and implementing them in our project.

We are also thankful to each team member for their every contribution in bringing all this together.

We are thankful to Airbnb for making their public dataset available and to AWS developers for their detailed guides on the various services.

References

[1] AWS Redshift. (2013). AWS Redshift. https://docs.aws.amazon.com/redshift/latest/gsg/getting-started.html

[2] AWS Glue. (2017). ETL Tool- https://docs.aws.amazon.com/glue/latest/dg/populate-datacatalog.html

[3] J.M.P. Martinez, R.B. Llavori, M.J.A. Cabo, and T.B. Pedersen, "Integrating Data Warehouses with Web Data: A Survey," *IEEE Trans. Knowledge and Data Eng.*, preprint, 21 Dec. 2007, doi:10.1109/TKDE.2007.190746. (PrePrint)

[4] Gupta, Anurag & Agarwal, Deepak & Tan, Derek & Kulesza, Jakub & Pathak, Rahul & Stefani, Stefano & Srinivasan, Vidhya. (2015). Amazon Redshift and the Case for Simpler Data Warehouses. 1917-1923. 10.1145/2723372.2742795.

[5] Inside Airbnb. 2021. *Inside Airbnb. Adding data to the debate.* [online] Available at: <http://insideairbnb.com/get-the-data.html> [Accessed 10 March 2021].

[6] Sans, A. and Domínguez, A., 2021. *13. Unravelling Airbnb: Urban Perspectives from Barcelona*. [online] De Gruyter. Available at: <https://www.degruyter.com/document/doi/10.21832/9781845415709-015/html> [Accessed 10 March 2021].

[7] Gábor, D., György, V., Kovalcsik, K. and Lajos, B., 2021. *A socio-economic analysis of Airbnb in New York City*. [online] Ceeol.com. Available at: <https://www.ceeol.com/search/article-detail?id=578374> [Accessed 10 March 2021].

[8] W. Min and L. Lu, "*Who Wants to Live Like a Local? An Analysis of Determinants of Consumers' Intention to Choose AirBNB*," 2017 International Conference on Management Science and Engineering (ICMSE), Nomi, Japan, 2017, pp. 642-651, DOI: 10.1109/ICMSE.2017.8574467.

[9] M. Abdar, K. Lai and N. Y. Yen, "*Crowd Preference Mining and Analysis Based on Regional Characteristics on Airbnb*," 2017 3rd IEEE International Conference on Cybernetics (CYBCONF), Exeter, 2017, pp. 1-6, DOI: 10.1109/CYBConf.2017.7985771.

[10] Cheng, M. and Jin, X., 2021. *What do Airbnb users care about? An analysis of online review comments*. [online] Available at: <https://www.sciencedirect.com/science/article/pii/S0278431917307491> [Accessed 11 March 2021].

APPENDIX:

Term Project Rubric criteria	**points**	**Included in project**
Presentation Skills Includes time management	5	
Significance to the real world	3	Current business case analysis
Code Walkthrough	4	
Report Format, completeness, language, plagiarism, whether Turnitin could process it (no unnecessary screenshots), etc	5	
Version Control Use of Git / GitHub or equivalent; must be publicly accessible	3	We used GitHub as the code repository https://github.com/vraj1231/Airbnb_Data225_project
Discussion / Q&A	5	
Lessons learned Included in the report and presentation? How substantial and unique are they?	5	
Innovation	5	
Teamwork	5	All of us worked together on this project.
Technical difficulty	4	We solved all the technical difficulties which we mentioned in the report
Practiced pair programming? See https://en.wikipedia.org/wiki/Pair_pr	2	Yes. We had zoom calls and practiced pair programming.

ogrammin-gLinksLinksLinksLinksLinksLinksLinksLinks to an external site. to an external site. to an external site. to an external site. to an external site. to an external site. to an external site. to an external site.		Because of the pandemic, we had to practice it online.
Practiced agile / scrum (1-week sprints)? Submit evidence on Canvas - meeting minutes, other artifacts	3	Yes. We had uploaded the zoom links in the discussions tab under Group 8.
Used Grammarly / other tools for language? Grammarly's free version is sufficient; can use other tools as well. Submit report screenshot on Canvas.	2	Yes. we checked the content using Grammarly.
Elevator pitch video Create and upload a video to YouTube or other providers of your choice describing your project in 5 minutes or less. See for instance: https://www.youtube.com/watch?v=XJfYGZI4QpoLinksLinksLinks to an external site. to an external site. to an external site.	3	https://www.youtube.com/watch?v=Cx6ht8a8UTY
Slides	5	
Demo	5	
Used unique tools E.g.: LaTeX for writing report (submit .tex that is not generated from another format such as .docx; generating from .lyx and similar LaTeX editor outputs is fine) Unique features of Prezi or PowerPoint, etc	5	
Performed substantial analysis using database techniques. The project must include an analytics component	3	We used SQL queries, MS Excel, QuickSight
Used a new database or data warehouse tool not covered in the HW or class	5	We used Redshift as the data warehouse
Used appropriate data models	5	yes
Used ETL Tool	3	AWS Glue is used as an ETL tool
Data Cleansing	2	We removed null values, duplicates, did case conversion as a part of the Glue transform job
Demonstrated how Analytics support business decisions	2	Included analysis
Used NoSQL database Idea is to exercise as many topics from the course as possible	3	We used AWS DynamoDB to store the geojson data.
Used RDBMS Idea is to exercise as many topics from the course as possible	2	Yes. Redshift operates on PostgresSQL.
Used Datawarehouse Idea is to exercise as many topics from the course as possible	3	AWS Redshift is used as the data warehouse
Includes DB Connectivity / API calls Possibly using Python	3	Yes. We used a Python program to connect to DynamoDB and generate tables from the script.